

Structural Properties of Bayesian Bandits with Exponential Family Distributions

Yaming Yu

Department of Statistics
University of California
Irvine, CA 92697, USA
yamingy@uci.edu

Abstract

We study a bandit problem where observations from each arm have an exponential family distribution and different arms are assigned independent conjugate priors. At each of n stages, one arm is to be selected based on past observations. The goal is to find a strategy that maximizes the expected discounted sum of the n observations. Two structural results hold in broad generality: (i) for a fixed prior weight, an arm becomes more desirable as its prior mean increases; (ii) for a fixed prior mean, an arm becomes more desirable as its prior weight decreases. These generalize and unify several results in the literature concerning specific problems including Bernoulli and normal bandits. The second result captures an aspect of the exploration-exploitation dilemma in precise terms: given the same immediate payoff, the less one knows about an arm, the more desirable it becomes because there remains more information to be gained when selecting that arm. For Bernoulli and normal bandits we also obtain extensions to nonconjugate priors.

Keywords: Bernoulli bandits; convex order; log-concavity; optimal stopping; sequential decision; two-armed bandits.

MSC 2010: Primary 62L05, 62C10; Secondary 62L15, 60E15.

1 Introduction

At each of n stages, an experimenter must take an observation from one of two stochastic processes (arms). Let us adopt the Bayesian framework and assume that the experimenter's belief about an unknown arm is updated according to Bayes Theorem after each observation. A strategy specifies which process to select at each stage. The objective is to maximize the

expected payoff, $\sum_{i=1}^n a_i Z_i$, where Z_i is the observation at stage i and $A_n \equiv (a_1, a_2, \dots, a_n)$ is a discount sequence satisfying $a_i \geq 0$ and $\sum_{i=1}^n a_i > 0$. A strategy is optimal if it achieves the maximum expected payoff. This is a finite-horizon two-armed bandit (Berry and Fristedt 1985), a classical problem in sequential decision theory.

Bernoulli bandits, where each arm generates binary observations, are important as a model for clinical trials, and have received considerable attention (Berry 1972; Berry and Fristedt 1985). Others such as normal (Chernoff 1968; Chernoff and Petkau 1986; Yao 2006) and Dirichlet bandits (Clayton and Berry 1985; Yu 2011) have also been extensively studied. Bandit problems exhibit a well-known exploration-exploitation tradeoff. Simply maximizing the immediate payoff is usually not an optimal strategy; one must allow for exploring an unknown arm for higher payoff later on. From a Bayesian perspective, the optimal strategy is easily specified through backward induction, although its computation can be nontrivial. If the discount sequence is geometric, then the problem reduces to several one-armed bandits (Gittins and Jones 1974; Gittins 1979; Whittle 1980; Kaspri and Mandelbaum 1998) and the optimal strategy is to choose an arm with the highest dynamic allocation index, or Gittins index. Optimal strategies for general discount sequences are less tractable.

The Gittins index possesses intriguing monotonicity properties with respect to prior specifications. For example, Gittins and Wang (1992) show that the Gittins index decreases in $\tau > 0$ for some special bandit arms: a Bernoulli arm whose unknown parameter has a $\text{Beta}(\tau s, \tau(1-s))$ prior ($0 < s < 1$), or a normal arm whose unknown mean has a $N(\mu, 1/\tau)$ prior ($\mu \in \mathbf{R}$). In both cases τ is naturally interpreted as the amount of prior information. Such monotonicity results therefore capture an aspect of the exploration-exploitation dilemma in precise terms: given the same immediate payoff, the less one knows about an arm, the more desirable it becomes since there is more room for exploration. In the literature, however, this monotonicity is usually derived for one-armed bandits and on a case-by-case basis. This paper aims to obtain more general results in a unified framework.

The Bernoulli and normal bandits can be regarded as special cases of a general bandit where observations from each arm have an exponential family distribution. Assume each arm is assigned an independent conjugate prior, which is characterized by a prior mean and a prior weight. The prior mean specifies the immediate payoff of an arm, whereas the prior weight reflects the associated uncertainty. For such problems we show that: (i) for fixed prior weight, the maximum expected payoff increases as the prior mean for any arm increases; (ii) for fixed prior mean, the maximum expected payoff increases as the prior weight for any arm decreases. These

generalize and unify several results in the literature concerning specific distributions. Similar techniques yield parallel results for Dirichlet bandits, which do not fit in the one-parameter exponential family framework (Clayton and Berry 1985; Chattopadhyay 1994; Yu 2011).

The rest of the paper is organized as follows. After setting up the exponential family framework and introducing a few notions of stochastic ordering in Section 2, we present basic structural results such as a stay-on-a-winner rule in Section 3. Section 4 contains the main results, including monotonicity of the value function with respect to prior weights. Section 5 applies the results in Section 4 to one-armed bandits. In particular, we show that the break-even value decreases as the prior weight of the unknown arm increases. In Sections 6 and 7 we extend the monotonicity results to nonconjugate priors for Bernoulli and normal bandits, respectively. Section 8 concludes with a brief discussion on an open problem.

2 Preliminaries

Let ν be a σ -finite measure on \mathbf{R} that is not a point mass. Denote

$$\psi(\theta) = \log \int e^{\theta x} d\nu(x), \quad \theta \in \Theta,$$

where Θ is the natural parameter space defined as the set of $\theta \in \mathbf{R}$ such that $\psi(\theta)$ is finite. We assume that Θ has a non-empty interior. Suppose that given θ_i , observations from arm i are independent and identically distributed (i.i.d.) according to the density (relative to ν)

$$f(x|\theta_i) = e^{\theta_i x - \psi(\theta_i)}. \quad (1)$$

Let us assume independent conjugate priors on θ_i , $i = 1, 2$, with Lebesgue density

$$f(\theta_i|\gamma_i, \tau_i) \propto e^{\theta_i \gamma_i - \tau_i \psi(\theta_i)}, \quad \theta_i \in \Theta. \quad (2)$$

Let \mathcal{K} denote the smallest open interval such that ν assigns no mass outside of the closure $\bar{\mathcal{K}}$. To ensure that the priors are proper, we require $\tau_i > 0$ and $\gamma_i/\tau_i \in \mathcal{K}$ (Brown 1986, Chapter 4). As usual τ_i is regarded as the “prior sample size” and γ_i the “prior sum of observations”. We refer to (2) as the (γ_i, τ_i) prior and call this two-armed bandit with discount sequence A_n the $(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ bandit. Its value (i.e., maximum expected payoff) is denoted by $V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$.

This framework unifies several well-studied bandit reward structures: (i) Bernoulli rewards whose unknown parameter has a $\text{Beta}(\gamma, \tau - \gamma)$ prior; (ii) normal rewards whose unknown

mean has a $N(\gamma/\tau, 1/\tau)$ prior; (iii) exponential rewards whose unknown rate parameter has a $\text{Gamma}(\tau + 1, \gamma)$ prior; (iv) Poisson rewards whose unknown rate parameter has a $\text{Gamma}(\gamma, \tau)$ prior. Extensions to general priors for (i) and (ii) are considered in Sections 6 and 7, respectively.

Let $V^i(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ be the expected payoff when selecting arm i initially and using an optimal strategy thereafter. Then

$$V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = \max \{V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n), V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)\}, \quad (3)$$

and it is optimal to start with the arm whose V^i is larger. Suppose arm 1 is selected, resulting in an observation X . By conjugacy, the posterior for θ_1 is again of the form of (2) with $(\gamma_1 + X, \tau_1 + 1)$ in place of (γ_1, τ_1) . Thus we have

$$V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = a_1 \mu_1 + E[V(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \gamma_1, \tau_1], \quad (4)$$

$$V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = a_1 \mu_2 + E[V(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2], \quad (5)$$

where $A_n^1 = (a_2, a_3, \dots, a_n)$ and μ_i denotes the expected value of an observation from arm i under the (γ_i, τ_i) prior. This μ_i is simply $\mu_i = \gamma_i/\tau_i$, which we refer to as the prior mean. In $E[g(X) | \gamma_1, \tau_1]$, we use X to denote a generic observation from arm 1 under the (γ_1, τ_1) prior; similarly for Y . That is, the density of X relative to ν is

$$f(x) \propto \int_{\Theta} e^{\theta(\gamma_1 + x) - (\tau_1 + 1)\psi(\theta)} d\theta. \quad (6)$$

The dynamic programming equations (3)–(5) are crucial for both theoretical analysis and numerical computation of the optimal strategy.

A key tool in our derivation is the notion of stochastic ordering (Müller and Stoyan 2002; Shaked and Shanthikumar 2007). We shall use the usual stochastic order \leq_{st} , the convex order \leq_{cx} , the likelihood ratio order \leq_{lr} , and the relative log-concavity order \leq_{lc} . For random variables Z_1 and Z_2 taking values on \mathbf{R} , we write $Z_1 \leq_{\text{st}} Z_2$ (respectively, $Z_1 \leq_{\text{cx}} Z_2$), if $E\phi(Z_1) \leq E\phi(Z_2)$ for every increasing (respectively, convex) function ϕ such that the expectations exist. If $Z_1 \leq_{\text{st}} Z_2$ then we also say Z_2 is to the right of Z_1 . If Z_1 and Z_2 have densities $f_1(z)$ and $f_2(z)$ respectively, supported on the same interval, then we write $Z_1 \leq_{\text{lr}} Z_2$ (respectively, $Z_1 \leq_{\text{lc}} Z_2$) if $\log(f_1(z)/f_2(z))$ is decreasing (respectively, concave) in z . For example, the (γ, τ) prior increases in the likelihood ratio order as γ increases, and decreases in the relative log-concavity order as τ increases. (We use \leq_{lr} , \leq_{st} , \leq_{lc} and \leq_{cx} with densities as well as random variables.) Useful properties include the implication $\leq_{\text{lr}} \implies \leq_{\text{st}}$. Assuming equal means, it also holds that \leq_{lc} implies \leq_{cx} . Intuitively, the relative log-concavity order compares the amount of information

as it is defined through curvatures of the log density functions. Both \leq_{lr} and \leq_{lc} are preserved under the prior-to-posterior updating, which makes them ideal for studying structural properties in bandit problems. The log-concavity order is also useful in other seemingly unrelated contexts (Whitt 1985; Yu 2009a, 2009b, 2010).

3 Stay-on-a-winner

This section derives a basic monotonicity property of the optimal strategy: as the observation from an arm becomes larger, the inclination to pull that arm again also increases. Under suitable conditions we prove a generalized stay-on-a-winner rule, which is a natural extension of the results for Bernoulli bandits (Bradt, Johnson and Karlin 1956; Berry 1972; Berry and Fristedt 1985).

Let us define the advantage of arm 1 over arm 2 as

$$\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) - V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n).$$

Define $\Delta^+ = \max\{\Delta, 0\}$ and $\Delta^- = \min\{\Delta, 0\}$. By considering the initial two pulls one can show (Berry 1972)

$$\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = (a_1 - a_2) \left(\frac{\gamma_1}{\tau_1} - \frac{\gamma_2}{\tau_2} \right) \quad (7)$$

$$+ E [\Delta^+(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \gamma_1, \tau_1] \quad (8)$$

$$+ E [\Delta^-(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2]. \quad (9)$$

Proposition 1 states that as the prior mean of arm 1 increases, so does the advantage of arm 1 over arm 2, assuming A_n is decreasing. This can be extended to non-conjugate priors. Specifically, Δ increases as the prior for arm 1 becomes larger in the likelihood ratio order. Extensions to general Markov decision problems are also possible (Rieder and Wagner 1991). We provide a complete proof which serves as an introduction to the derivation of the main results in Section 4.

Proposition 1. *Suppose A_n is decreasing. Then $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ increases in γ_1 .*

Proof. The $n = 1$ case is easy. Let us use induction for $n \geq 2$. In view of (7)–(9), we only need to show that

$$E [\Delta^+(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \gamma_1, \tau_1] \quad \text{and} \quad (10)$$

$$E [\Delta^-(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2] \quad (11)$$

both increase in γ_1 . Monotonicity of (11) follows from the induction hypothesis. To handle (10), let us consider $\gamma_1 < \tilde{\gamma}_1$. Let θ_1 and $\tilde{\theta}_1$ have the (γ_1, τ_1) and $(\tilde{\gamma}_1, \tau_1)$ priors respectively. Let $g(x)$ (respectively, $\tilde{g}(x)$) be the marginal density of X if it is drawn according to (1) given θ_1 (respectively, $\tilde{\theta}_1$). Note that $\theta_1 \leq_{\text{lr}} \tilde{\theta}_1$. In view of (6), we know that $g \leq_{\text{lr}} \tilde{g}$ by total positivity considerations (Karlin 1968, Chapter 3). It follows that $g \leq_{\text{st}} \tilde{g}$. By the induction hypothesis,

$$\phi(x) \equiv \Delta^+(x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1)$$

increases in x . Thus

$$\begin{aligned} E[\phi(\gamma_1 + X) | \gamma_1, \tau_1] &\leq E[\phi(\tilde{\gamma}_1 + X) | \gamma_1, \tau_1] \\ &\leq E[\phi(\tilde{\gamma}_1 + X) | \tilde{\gamma}_1, \tau_1], \end{aligned} \quad (12)$$

where (12) holds because $g \leq_{\text{st}} \tilde{g}$. Hence (10) increases in γ_1 . \square

Corollary 1. *Suppose A_n is a decreasing sequence, and an observation x is taken from arm 1 initially. Then, at the second stage, either arm 1 is optimal for all x , or arm 2 is optimal for all x , or there exists some $x_* \in \mathcal{K}$ such that arm 1 is optimal if $x \geq x_*$ and arm 2 is optimal if $x \leq x_*$.*

Proof. We can show that $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1)$ is continuous in x . (One method is to use the convexity result of Proposition 2 in Section 4.) The claim then follows from Proposition 1. \square

The next result, Theorem 1, is a generalized stay-on-a-winner rule: under suitable conditions if an arm is optimal initially then it continues to be optimal at the next stage provided that the initial observation from that arm is large enough.

Theorem 1. *Assume A_n is decreasing, $n \geq 2$, and either (i) $a_1 = a_2$ or (ii) $\gamma_1/\tau_1 \leq \gamma_2/\tau_2$ holds. Assume $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \geq 0$, i.e., arm 1 is optimal initially. Then $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) \geq 0$ for sufficiently large $x \in \bar{\mathcal{K}}$.*

Proof. We may assume $a_i > 0$ for all $i \leq n$. Let U be the upper end point of \mathcal{K} . If $U = \infty$, then using (7)–(9), it is easy to show by induction that $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) > 0$ for sufficiently large x . That is, the claim holds even without assuming that arm 1 is optimal initially. Assume $U < \infty$ and $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \geq 0$. By (7)–(9) we have

$$0 \leq E[\Delta^+(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \gamma_1, \tau_1] + E[\Delta^-(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2]. \quad (13)$$

Suppose the claim does not hold, i.e., $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) < 0$ for all $x \in \bar{\mathcal{K}}$. In particular,

$$\Delta(\gamma_1 + U, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) < 0. \quad (14)$$

Then it is necessary that both expectations in (13) are zero. That is,

$$\Delta(\gamma_1, \tau_1; \gamma_2 + y, \tau_2 + 1; A_n^1) \geq 0 \quad \text{for all } y \in \mathcal{K}.$$

By continuity, $\Delta(\gamma_1, \tau_1; \gamma_2 + U, \tau_2 + 1; A_n^1) \geq 0$. However, the $(\gamma_1 + U, \tau_1 + 1)$ prior is larger than the (γ_1, τ_1) prior in the likelihood ratio order. The argument of Proposition 1 yields

$$\begin{aligned} \Delta(\gamma_1 + U, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) &\geq \Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n^1) \\ &\geq \Delta(\gamma_1, \tau_1; \gamma_2 + U, \tau_2 + 1; A_n^1) \geq 0, \end{aligned}$$

which contradicts (14). □

4 Monotonicity

Proposition 2 shows that the maximum expected payoff is an increasing and convex function of the prior mean of any arm. The convexity will be useful in proving Theorem 2 concerning monotonicity with respect to the prior weight.

Proposition 2. *$V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ is increasing and convex in each of γ_i , $i = 1, 2$.*

Proof. Monotonicity holds by the same argument that proves Proposition 1. Let us focus on the convexity with respect to γ_1 . The $n = 1$ case is easy. For $n \geq 2$ we use induction. Note that by (3)–(5) it suffices to show that both

$$E [V(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \gamma_1, \tau_1] \quad \text{and} \quad (15)$$

$$E [V(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2] \quad (16)$$

are convex in γ_1 . The claim for (16) follows from the induction hypothesis. To deal with (15), suppose $\gamma_1 < \tilde{\gamma}_1$. Denote the marginal of X when the prior on θ is (γ_1, τ_1) (respectively, $(\tilde{\gamma}_1, \tau_1)$) by g (respectively, \tilde{g}). Then $g \leq_{\text{st}} \tilde{g}$ as in the proof of Proposition 1. By the induction hypothesis,

$$\phi(x) \equiv V(x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1)$$

is convex in x . Moreover,

$$\begin{aligned} E [\phi(\gamma_1 + X) | \gamma_1, \tau_1] &- E \left[\phi \left(\frac{\gamma_1 + \tilde{\gamma}_1}{2} + X \right) \middle| \gamma_1, \tau_1 \right] \\ &\geq E [\eta(X) | \gamma_1, \tau_1] \end{aligned} \tag{17}$$

$$\geq E [\eta(X) | \tilde{\gamma}_1, \tau_1] \tag{18}$$

where

$$\eta(x) \equiv \phi \left(\frac{\gamma_1 + \tilde{\gamma}_1}{2} + x \right) - \phi(\tilde{\gamma}_1 + x).$$

The inequality (17) holds because ϕ is convex; (18) holds because η is decreasing and $g \leq_{\text{st}} \tilde{g}$. Rearranging we get

$$E [\phi(\gamma_1 + X) | \gamma_1, \tau_1] + E [\phi(\tilde{\gamma}_1 + X) | \tilde{\gamma}_1, \tau_1] \geq 2E\phi \left(\frac{\gamma_1 + \tilde{\gamma}_1}{2} + X^* \right)$$

where X^* has the following distribution. Given θ , X^* is distributed according to (1); the prior on θ is a half-half mixture of (γ_1, τ_1) and $(\tilde{\gamma}_1, \tau_1)$. Denote this mixture density by $h^*(\theta)$, and the $((\gamma_1 + \tilde{\gamma}_1)/2, \tau_1)$ prior density by $h(\theta)$. Then $h(\theta) \leq_{\text{lc}} h^*(\theta)$, because log-convexity is closed under mixtures (Marshall and Olkin 1979). Consider the difference between the marginal densities

$$D(x) \equiv \int_{\Theta} e^{x\theta - \psi(\theta)} [h(\theta) - h^*(\theta)] d\theta.$$

Relative log-concavity implies that, as θ traverses Θ , $h(\theta) - h^*(\theta)$ changes signs at most twice and, in the case of two changes, the sign sequence is $-, +, -$. By the variation-diminishing properties of the Laplace transform (Karlin 1968, Chapter 5), $D(x)$ has at most two changes of sign, and in the case of two changes, the sign sequence is $-, +, -$. Note that, when the prior is either h or h^* , the marginal mean of X is the same, namely $(\gamma_1 + \tilde{\gamma}_1)/(2\tau_1)$. Hence it is not possible for $D(x)$ to change signs exactly once. Unless $D(x) \equiv 0$, its sign sequence must be $-, +, -$. It follows that the marginal distribution of X becomes larger in the convex order when $\tilde{h}(\theta)$ replaces $h(\theta)$ as the prior for θ (see, e.g., Yu 2010, Lemma 1). Using the convexity of ϕ again, we obtain

$$E\phi \left(\frac{\gamma_1 + \tilde{\gamma}_1}{2} + X^* \right) \geq E \left[\phi \left(\frac{\gamma_1 + \tilde{\gamma}_1}{2} + X \right) \middle| \frac{\gamma_1 + \tilde{\gamma}_1}{2}, \tau_1 \right].$$

It follows that $E[\phi(\gamma_1 + X) | \gamma_1, \tau_1]$, i.e., (15), is convex in γ_1 , as required. \square

Our main result, Theorem 2, shows that the value of the bandit decreases as the prior weight of an arm increases. That is, given the same immediate payoff, an arm becomes less desirable as the amount of information about it increases.

Theorem 2. $V(c\gamma_1, c\tau_1; \gamma_2, \tau_2; A_n)$ decreases in $c \in (0, \infty)$.

Proof. Let us use induction on n . The $n = 1$ case is easy. Suppose $n \geq 2$. In view of (3)–(5), we only need to show that

$$E[V(c\gamma_1 + X, c\tau_1 + 1; \gamma_2, \tau_2; A_n^1) | c\gamma_1, c\tau_1] \quad \text{and} \quad (19)$$

$$E[V(c\gamma_1, c\tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1) | \gamma_2, \tau_2] \quad (20)$$

both decrease in c . By the induction hypothesis, (20) decreases in c . To deal with (19), suppose $0 < c < \tilde{c}$ and denote $\xi = (c\tau_1 + 1)/(\tilde{c}\tau_1 + 1)$. We get

$$\begin{aligned} & E[V(c\gamma_1 + X, c\tau_1 + 1; \gamma_2, \tau_2; A_n^1) | c\gamma_1, c\tau_1] \\ & \geq E[V(\xi(\tilde{c}\gamma_1 + X), c\tau_1 + 1; \gamma_2, \tau_2; A_n^1) | c\gamma_1, c\tau_1] \end{aligned} \quad (21)$$

$$\geq E[V(\tilde{c}\gamma_1 + X, \tilde{c}\tau_1 + 1; \gamma_2, \tau_2; A_n^1) | c\gamma_1, c\tau_1] \quad (22)$$

$$\geq E[V(\tilde{c}\gamma_1 + X, \tilde{c}\tau_1 + 1; \gamma_2, \tau_2; A_n^1) | \tilde{c}\gamma_1, \tilde{c}\tau_1]. \quad (23)$$

The inequality (21) holds by the convexity of V as shown by Proposition 2, noting

$$\xi(\tilde{c}\gamma_1 + X) \leq_{\text{cx}} c\gamma_1 + X$$

(see Lemma 3 in Section 7, or Shaked and Shanthikumar 2007, Theorem 3.A.18). The inequality (22) holds by the induction hypothesis, as $\xi < 1$. The inequality (23) holds by an argument similar to the proof of Proposition 2. Specifically, the prior $(\tilde{c}\gamma_1, \tilde{c}\tau_1)$ is log-concave relative to $(c\gamma_1, c\tau_1)$. Thus the marginal of X increases in the convex order if $(c\gamma_1, c\tau_1)$ replaces $(\tilde{c}\gamma_1, \tilde{c}\tau_1)$ as the prior on θ (the mean of X remains constant). Overall (19) decreases in c , as required. \square

Remark. Proposition 2 and Theorem 2 extend naturally to bandits with more than two arms. We present the two-armed version for simplicity. The discount sequence A_n is only required to be nonnegative. By approximation, this can be further extended to the infinite-horizon case assuming $\sum_{i=1}^{\infty} a_i < \infty$.

5 The one-armed case

This section considers the one-armed case assuming that arm 2 yields a constant payoff λ at each pull. We shall abuse the notation by calling this a $(\gamma, \tau; \lambda; A_n)$ bandit, where we drop the subscripts on γ_1 and τ_1 for convenience. Results in Section 4 are applied to derive monotonicity properties of the break-even value in this case. It is also shown (Proposition 3) that if both

arms are optimal initially, then an observation from arm 1 that is less than its prior mean would make arm 2 optimal thereafter.

A discount sequence $A_n = (a_1, a_2, \dots)$ is called *regular* if, letting $b_j = \sum_{i \geq j} a_i$, we have $b_{j+1}^2 \geq b_j b_{j+2}$ for all $j \geq 1$ (Berry and Fristedt 1979). For regular discount sequences, our one-armed bandit is an optimal stopping problem, i.e., if at any stage the known arm becomes optimal then it remains optimal in all subsequent stages. Moreover, if A_n is regular and $a_1 > 0$, then there exists a break-even value $\Lambda(\gamma, \tau; A_n)$ for the $(\gamma, \tau; \lambda; A_n)$ bandit, such that arm 1 is optimal initially if and only if $\lambda \leq \Lambda(\gamma, \tau; A_n)$ and arm 2 is optimal initially if and only if $\lambda \geq \Lambda(\gamma, \tau; A_n)$. For infinite-horizon geometric discounting, this break-even value is also known as the dynamic allocation index or Gittins index (Gittins and Jones 1974). The following result holds by the optimal stopping characterization.

Lemma 1. *If A_n is regular and $a_1 > 0$, then $\Lambda(\gamma, \tau; A_n)$ is the smallest λ such that*

$$V(\gamma, \tau; \lambda; A_n) \leq \lambda \sum_{i=1}^n a_i.$$

Corollary 2 summarizes some monotonicity properties of $\Lambda(\gamma, \tau; A_n)$. It extends to infinite-horizon regular discounting. As special cases we recover the results of Gittins and Wang (1992) on Bernoulli and normal bandits with geometric discounting; see also Yao (2006).

Corollary 2. *If A_n is regular and $a_1 > 0$, then $\Lambda(c\gamma, c\tau; A_n)$ decreases in $c > 0$ and strictly increases in γ .*

Proof. Monotonicity in c follows from Theorem 2 and Lemma 1. Monotonicity in γ follows from Proposition 2 and Lemma 1. To show strict monotonicity, let us set $c = 1$ and assume that $\gamma, \tilde{\gamma}$ satisfy $\gamma < \tilde{\gamma}$ and

$$\Lambda(\gamma, \tau; A_n) = \Lambda(\tilde{\gamma}, \tau; A_n) \equiv \lambda_*.$$

Then, as in the proof of Proposition 1, we get

$$\begin{aligned} \lambda_* \sum_{i=1}^n a_i &= a_1 \frac{\gamma}{\tau} + E[V(\gamma + X, \tau + 1; \lambda_*; A_n^1) | \gamma, \tau] \\ &< a_1 \frac{\tilde{\gamma}}{\tau} + E[V(\gamma + X, \tau + 1; \lambda_*; A_n^1) | \gamma, \tau] \\ &\leq a_1 \frac{\tilde{\gamma}}{\tau} + E[V(\tilde{\gamma} + X, \tau + 1; \lambda_*; A_n^1) | \tilde{\gamma}, \tau] \\ &= \lambda_* \sum_{i=1}^n a_i, \end{aligned}$$

which is a contradiction. □

For a regular and positive discount sequence A_n , Proposition 3 shows that there exists a break-even observation $b(\gamma, \tau; A_n)$ for the $(\gamma, \tau; \lambda; A_n)$ bandit such that if both arms are optimal initially, and an observation x is taken from arm 1, then arm 1 remains optimal if $x \geq b(\gamma, \tau; A_n)$ and arm 2 becomes optimal if $x \leq b(\gamma, \tau; A_n)$. Moreover, this break-even observation is no smaller than γ/τ , the prior mean.

Proposition 3. *Suppose A_n is regular, $n \geq 2$, and $a_1, a_2 > 0$. Then there exists a unique $b(\gamma, \tau; A_n) \in \mathcal{K}$ such that $b(\gamma, \tau; A_n) \geq \gamma/\tau$ and*

$$\Lambda(\gamma, \tau; A_n) \geq \Lambda(\gamma + x, \tau + 1; A_n^1), \quad \text{if } x \leq b(\gamma, \tau; A_n); \quad (24)$$

$$\Lambda(\gamma, \tau; A_n) \leq \Lambda(\gamma + x, \tau + 1; A_n^1), \quad \text{if } x \geq b(\gamma, \tau; A_n). \quad (25)$$

To prove Proposition 3 we need a continuity lemma. Its proof, taken from Clayton and Berry (1985), is included for completeness.

Lemma 2. *Suppose A_n is regular and $a_1 > 0$. Then $\Lambda(\gamma, \tau; A_n)$ is continuous in γ .*

Proof. Fix γ_0 and note that $\lambda = \Lambda(\gamma, \tau; A_n)$ is the unique root of

$$V^1(\gamma, \tau; \lambda; A_n) - V^2(\gamma, \tau; \lambda; A_n) = 0.$$

By continuity of V^1 and V^2 , we have

$$\begin{aligned} 0 &= \lim_{\gamma \uparrow \gamma_0} [V^1(\gamma, \tau; \Lambda(\gamma, \tau; A_n); A_n) - V^2(\gamma, \tau; \Lambda(\gamma, \tau; A_n); A_n)] \\ &= V^1(\gamma_0, \tau; \lim_{\gamma \uparrow \gamma_0} \Lambda(\gamma, \tau; A_n); A_n) - V^2(\gamma_0, \tau; \lim_{\gamma \uparrow \gamma_0} \Lambda(\gamma, \tau; A_n); A_n). \end{aligned}$$

By uniqueness of Λ , we have $\lim_{\gamma \uparrow \gamma_0} \Lambda(\gamma, \tau; A_n) = \Lambda(\gamma_0, \tau; A_n)$. Similarly, the limit holds when $\gamma \downarrow \gamma_0$. \square

Proof of Proposition 3. Let U be the upper end point of \mathcal{K} . If $U = \infty$ then $\Lambda(\gamma + x, \tau + 1; A_n^1) \rightarrow \infty$ as $x \rightarrow \infty$ (the expected payoff by always selecting arm 1 becomes arbitrarily large). If $U < \infty$ then we can show $\Lambda(\gamma + U, \tau + 1; A_n^1) > \Lambda(\gamma, \tau; A_n)$ as follows. Assume the contrary and consider the $(\gamma, \tau; \lambda_*; A_n)$ bandit with $\lambda_* = \Lambda(\gamma + U, \tau + 1; A_n^1)$. We have

$$\lambda_* \sum_{i=1}^n a_i \leq a_1 \frac{\gamma}{\tau} + E[V(\gamma + X, \tau + 1; \lambda_*; A_n^1) | \gamma, \tau].$$

Since $\gamma/\tau \in \mathcal{K}$ and \mathcal{K} is open, we have $\lambda_* \geq (\gamma + U)/(\tau + 1) > \gamma/\tau$. Thus

$$\begin{aligned} \lambda_* \sum_{i=2}^n a_i &< E[V(\gamma + X, \tau + 1; \lambda_*; A_n^1) | \gamma, \tau] \\ &\leq V(\gamma + U, \tau + 1; \lambda_*; A_n^1) \\ &= \lambda_* \sum_{i=2}^n a_i, \end{aligned}$$

which is a contradiction. We also have

$$\begin{aligned} \Lambda(\gamma, \tau; A_n) &\geq \Lambda(\gamma, \tau; A_n^1) \\ &\geq \Lambda(\gamma + \gamma/\tau, \tau + 1; A_n^1) \end{aligned}$$

where the first inequality holds by the optimal stopping characterization, and the second by Corollary 2.

By Lemma 2 and Corollary 2, $\Lambda(\gamma + x, \tau + 1; A_n^1)$ is continuous and strictly increasing in x . By the mean value theorem, there exists a unique $b(\gamma, \tau; A_n) \in [\gamma/\tau, U)$ such that (24) and (25) hold. \square

It is tempting to conjecture that $b(\gamma, \tau; A_n) \geq \Lambda(\gamma, \tau; A_n)$, which gives a tighter bound since $\Lambda(\gamma, \tau; A_n) \geq \gamma/\tau$. However, our methods are not yet strong enough to resolve this conjecture. Clayton and Berry (1985) conjectured and Yu (2011) proved an analogous bound for Dirichlet bandits.

6 Bernoulli bandits with general priors

As noted earlier, results based on likelihood ratio orders, such as those in Section 3, may extend to nonconjugate priors. This section shows that Theorem 2 can also be extended this way, at least in the Bernoulli case.

Given p_i , $i = 1, 2$, let us assume that observations from arm i are i.i.d. Bernoulli(p_i). Priors on p_i are independent with densities f_i with respect to a σ -finite measure G on $[0, 1]$. We shall denote the value of this Bernoulli bandit with discount sequence A_n by $V_B(f_1; f_2; A_n)$. Let $\mu(f)$ denote the mean of any prior f , i.e., $\mu(f) = \int_{[0,1]} pf(p) dG(p)$.

Theorem 3. *If $f_1 \leq_{lc} \tilde{f}_1$ and $\mu(f_1) = \mu(\tilde{f}_1)$, then $V_B(f_1; f_2; A_n) \leq V_B(\tilde{f}_1; f_2; A_n)$.*

Note that the Beta($c\alpha, c\beta$) prior ($c, \alpha, \beta > 0$) decreases in the relative log-concavity order as c increases. Theorem 3 therefore recovers the Bernoulli case of Theorem 2 for conjugate priors.

Let $\Lambda_B(f; A_n)$ denote the break-even value of a one armed Bernoulli bandit whose unknown arm has prior f . We obtain Corollary 3 as a consequence of Theorem 3 and Lemma 1.

Corollary 3. *Assume A_n is regular and $a_1 > 0$. If $f \leq_{\text{lc}} \tilde{f}$ and $\mu(f) = \mu(\tilde{f})$, then $\Lambda_B(f; A_n) \leq \Lambda_B(\tilde{f}; A_n)$.*

Herschhorn (1997) posed the problem of identifying a variability ordering between priors so that both V_B and Λ_B are monotonic with respect to it. Theorem 3 and Corollary 3 show that there is indeed such an ordering, namely the relative log-concavity order (assuming equal means). A conjecture of Herschhorn (1997) states that Corollary 3 holds under the weaker assumption $f \leq_{\text{cx}} \tilde{f}$. This conjecture remains open.

Proof of Theorem 3. The $n = 1$ case is easy. For $n \geq 2$ we use induction. The equations (3)–(5) become

$$\begin{aligned} V_B(f_1; f_2; A_n) &= \max\{V_B^1(f_1; f_2; A_n), V_B^2(f_1; f_2; A_n)\}; \\ V_B^1(f_1; f_2; A_n) &= \mu(f_1)(a_1 + V_B(\sigma f_1; f_2; A_n^1)) + (1 - \mu(f_1))V_B(\phi f_1; f_2; A_n^1); \\ V_B^2(f_1; f_2; A_n) &= \mu(f_2)(a_1 + V_B(f_1; \sigma f_2; A_n^1)) + (1 - \mu(f_2))V_B(f_1; \phi f_2; A_n^1). \end{aligned} \quad (26)$$

We use σf (respectively, ϕf) to denote the posterior density after observing one success (respectively, one failure). That is,

$$(\sigma f)(p) = \frac{f(p)p}{\mu(f)}; \quad (\phi f)(p) = \frac{f(p)(1-p)}{1-\mu(f)}.$$

Let us assume \tilde{f}_1 is nondegenerate. Because $f_1 \leq_{\text{lc}} \tilde{f}_1$ and $\mu(f_1) = \mu(\tilde{f}_1)$ we have $f_1 \leq_{\text{cx}} \tilde{f}_1$ (see, e.g., Yu 2010, Theorem 12). Thus

$$\mu(f_1)\mu(\sigma f_1) = \int_{[0,1]} p^2 f_1(p) dG(p) \leq \int_{[0,1]} p^2 \tilde{f}_1(p) dG(p) = \mu(\sigma \tilde{f}_1)\mu(\tilde{f}_1),$$

yielding $\mu(\sigma f_1) \leq \mu(\sigma \tilde{f}_1)$. Similarly, $\mu(\phi f_1) \geq \mu(\phi \tilde{f}_1)$. Define

$$\epsilon^* = \frac{\mu(\sigma \tilde{f}_1) - \mu(\sigma f_1)}{\mu(\sigma \tilde{f}_1) - \mu(\phi \tilde{f}_1)}; \quad \epsilon_* = \frac{\mu(\phi f_1) - \mu(\phi \tilde{f}_1)}{\mu(\sigma \tilde{f}_1) - \mu(\phi \tilde{f}_1)}.$$

Then $\epsilon^*, \epsilon_* \in [0, 1)$. Define

$$g^* = (1 - \epsilon^*)\sigma \tilde{f}_1 + \epsilon^*\phi \tilde{f}_1; \quad g_* = \epsilon_*\sigma \tilde{f}_1 + (1 - \epsilon_*)\phi \tilde{f}_1.$$

Convexity of V_B with respect to mixtures gives

$$\begin{aligned} V_B(g^*; f_2; A_n^1) &\leq (1 - \epsilon^*)V_B(\sigma \tilde{f}_1; f_2; A_n^1) + \epsilon^*V_B(\phi \tilde{f}_1; f_2; A_n^1); \\ V_B(g_*; f_2; A_n^1) &\leq \epsilon_*V_B(\sigma \tilde{f}_1; f_2; A_n^1) + (1 - \epsilon_*)V_B(\phi \tilde{f}_1; f_2; A_n^1). \end{aligned}$$

Noting $\mu(f_1)\epsilon^* = (1 - \mu(f_1))\epsilon_*$, we add $\mu(f_1)$ times the first inequality to $1 - \mu(f_1)$ times the second and get

$$\begin{aligned} & \mu(f_1)V_B(g^*; f_2; A_n^1) + (1 - \mu(f_1))V_B(g_*; f_2; A_n^1) \\ & \leq \mu(f_1)V_B(\sigma\tilde{f}_1; f_2; A_n^1) + (1 - \mu(f_1))V_B(\phi\tilde{f}_1; f_2; A_n^1). \end{aligned} \quad (27)$$

The density g^* is simply

$$g^*(p) = \left[\frac{p(1 - \epsilon^*)}{\mu(f_1)} + \frac{(1 - p)\epsilon^*}{1 - \mu(f_1)} \right] \tilde{f}_1(p).$$

It is easy to check $(1 - \epsilon^*)/\mu(f_1) \geq \epsilon^*/(1 - \mu(f_1))$, which leads to

$$\sigma f_1 \leq_{\text{lc}} \sigma\tilde{f}_1 \leq_{\text{lc}} g^*.$$

Moreover, σf_1 and g^* have the same mean. By the induction hypothesis, we have

$$V_B(\sigma f_1; f_2; A_n^1) \leq V_B(g^*; f_2; A_n^1). \quad (28)$$

Similarly,

$$V_B(\phi f_1; f_2; A_n^1) \leq V_B(g_*; f_2; A_n^1). \quad (29)$$

We combine (27)–(29) to get

$$\begin{aligned} & \mu(f_1)V_B(\sigma f_1; f_2; A_n^1) + (1 - \mu(f_1))V_B(\phi f_1; f_2; A_n^1) \\ & \leq \mu(f_1)V_B(\sigma\tilde{f}_1; f_2; A_n^1) + (1 - \mu(f_1))V_B(\phi\tilde{f}_1; f_2; A_n^1). \end{aligned}$$

Applying (26) then yields

$$V_B^1(f_1; f_2; A_n) \leq V_B^1(\tilde{f}_1; f_2; A_n).$$

The rest of the proof is standard. \square

Remark. Theorem 3 focuses on the parameter p . If we still require equal prior means for p , but impose the log-concavity order on $\theta = \log(p/(1 - p))$ rather than p , then V_B is ordered by virtually the same proof. This result is distinct from Theorem 3 because the relative log-concavity order is usually not preserved by monotone transformations.

7 Normal bandits with general priors

The main result of this section (Theorem 4) extends Theorem 2 to general priors for normal bandits. Similar to Theorem 3, Theorem 4 is based on the relative log-concavity order, although it is more restrictive because we only compare a general prior with a normal prior.

Given θ_i , $i = 1, 2$, let us assume that observations from arm i are i.i.d. $N(\theta_i, 1)$. Priors on θ_i are independent with Lebesgue densities f_i . We shall denote the value of this normal bandit with discount sequence A_n by $V_N(f_1; f_2; A_n)$. Denote the mean of any f by $\mu(f) = \int_{-\infty}^{\infty} \theta f(\theta) d\theta$.

Theorem 4. *Let $\tilde{f}_1 \equiv N(\alpha, 1/\tau)$.*

1. *If $f_1 \leq_{lc} \tilde{f}_1$ and $\mu(f_1) = \alpha$, then $V_N(f_1; f_2; A_n) \leq V_N(\tilde{f}_1; f_2; A_n)$.*
2. *If $\tilde{f}_1 \leq_{lc} f_1$ and $\mu(f_1) = \alpha$, then $V_N(\tilde{f}_1; f_2; A_n) \leq V_N(f_1; f_2; A_n)$.*

Let $\Lambda_N(f; A_n)$ denote the break-even value of a one-armed normal bandit with prior f for the mean of the unknown arm. We obtain Corollary 4 as a consequence of Theorem 4 and Lemma 1.

Corollary 4. *Assume A_n is regular and $a_1 > 0$. Define $\tilde{f} \equiv N(\alpha, 1/\tau)$.*

1. *If $f \leq_{lc} \tilde{f}$ and $\mu(f) = \alpha$, then $\Lambda_N(f; A_n) \leq \Lambda_N(\tilde{f}; A_n)$.*
2. *If $\tilde{f} \leq_{lc} f$ and $\mu(f) = \alpha$, then $\Lambda_N(\tilde{f}; A_n) \leq \Lambda_N(f; A_n)$.*

The condition $f \leq_{lc} N(\alpha, 1/\tau)$ is essentially $d^2 \log f(\theta)/d\theta^2 \leq -\tau$, which can be regarded as a strong form of information ordering. The appearance of \leq_{lc} is therefore especially intuitive in Theorem 4 and Corollary 4. It is an open problem whether Theorem 4 and Corollary 4 hold without assuming that one of the priors is normal.

The rest of this section proves Theorem 4. We need a technical result (Lemma 3) which may be of independent interest.

Lemma 3. *Let g be a differentiable function on \mathbf{R} . Assume X is a random variable satisfying $Eg(X) = EX$.*

1. *If $0 \leq g'(x) \leq 1$, $x \in \mathbf{R}$, then $g(X) \leq_{cx} X$.*
2. *If $g'(x) \geq 1$, $x \in \mathbf{R}$, then $X \leq_{cx} g(X)$.*

Proof. We prove Part 1 only. Part 2 follows from Part 1 by considering the inverse function of g . As $Eg(X) = EX$, one criterion for $g(X) \leq_{\text{cx}} X$ is

$$E \max\{0, g(X) - b\} \leq E \max\{0, X - b\}, \quad b \in \mathbf{R}. \quad (30)$$

See, e.g., Shaked and Shanthikumar (2007; Theorem 3.A.1). Let us assume $0 \leq g'(x) \leq c$ for some $0 < c < 1$. Otherwise we consider $cg(x)$ and let $c \uparrow 1$. As $g(x)$ is a contraction, it has a unique fixed point, say x_0 . Consider two cases.

Case (i): $b \geq x_0$. If $x \geq x_0$ then $g(x) - g(x_0) \leq x - x_0$, i.e., $g(x) \leq x$, and $\max\{0, g(x) - b\} \leq \max\{0, x - b\}$. If $x < x_0$ then $g(x) \leq g(x_0) = x_0$ and

$$\max\{0, g(x) - b\} \leq \max\{0, x_0 - b\} = 0 \leq \max\{0, x - b\}.$$

In either case $\max\{0, g(x) - b\} \leq \max\{0, x - b\}$, which implies (30).

Case (ii): $b < x_0$. Applying the argument of Case (i) to $\tilde{g}(x) \equiv -g(-x)$ and $\tilde{X} \equiv -X$ yields $E \max\{0, b - g(X)\} \leq E \max\{0, b - X\}$, which reduces to (30) because $Eg(X) = EX$. \square

Proof of Theorem 4. We only prove Part 1; the second part is similar. The $n = 1$ case is easy. For $n \geq 2$ we use induction. The equations (3)–(5) become

$$\begin{aligned} V_N(f_1; f_2; A_n) &= \max \{ V_N^1(f_1; f_2; A_n), V_N^2(f_1; f_2; A_n) \}; \\ V_N^1(f_1; f_2; A_n) &= a_1 \mu(f_1) + E [V_N(f_1^X; f_2; A_n^1) | \Phi f_1]; \\ V_N^2(f_1; f_2; A_n) &= a_1 \mu(f_2) + E [V_N(f_1; f_2^Y; A_n^1) | \Phi f_2]. \end{aligned} \quad (31)$$

We denote the posterior $f_1^x(\theta) \propto f_1(\theta) \exp[-(x - \theta)^2/2]$; similarly for f_2^y . In $E[g(X) | \Phi f]$, the density of X , denoted by Φf , is the convolution of f with the standard normal. (Note the difference from the notation in Section 2.) Let $m(x; f)$ denote the posterior mean of θ when x is observed and the prior is f , i.e., $m(x; f) = \int_{-\infty}^{\infty} \theta f^x(\theta) d\theta$. Direct calculation yields

$$\frac{dm(x; f)}{dx} = \text{Var}(\theta | f^x). \quad (32)$$

That is, the derivative of $m(x; f)$ is simply the posterior variance of θ .

Suppose $f_1 \leq_{\text{lc}} \tilde{f}_1 \equiv N(\alpha, 1/\tau)$ and $\mu(f_1) = \alpha$. Then

$$f_1^x \leq_{\text{lc}} N \left(m(x; f_1), \frac{1}{\tau + 1} \right). \quad (33)$$

It can be shown that (i) if X is distributed as Φf_1 , then $m(X; f_1) \leq_{\text{cx}} (X + \tau\alpha)/(\tau + 1)$; (ii) Φf_1 is smaller than $\Phi \tilde{f}_1 \equiv N(\alpha, 1 + 1/\tau)$ in the convex order. To prove (i), note that (33) holds

with \leq_{lc} replaced by \leq_{cx} as the two sides have equal means. By (32) we have

$$0 \leq \frac{dm(x; f_1)}{dx} \leq \frac{1}{\tau + 1}, \quad x \in \mathbf{R}.$$

If X is distributed as Φf_1 then both $(X + \tau\alpha)/(\tau + 1)$ and $m(X; f_1)$ have mean $\mu(f_1) = \alpha$. Thus claim (i) holds by Lemma 3. Claim (ii) holds because $f_1 \leq_{\text{cx}} \tilde{f}_1$ and the convex order is closed under convolution.

We have

$$\begin{aligned} E [V_N(f_1^X; f_2; A_n^1) | \Phi f_1] &\leq E \left[V_N \left(N \left(m(X; f_1), \frac{1}{\tau + 1} \right); f_2; A_n^1 \right) \middle| \Phi f_1 \right] \\ &\leq E \left[V_N \left(\tilde{f}_1^X; f_2; A_n^1 \right) \middle| \Phi f_1 \right] \\ &\leq E \left[V_N \left(\tilde{f}_1^X; f_2; A_n^1 \right) \middle| \Phi \tilde{f}_1 \right], \end{aligned}$$

where the first inequality holds by (33) and the induction hypothesis, the second by claim (i), noting

$$\tilde{f}_1^X = N \left(\frac{X + \tau\alpha}{\tau + 1}, \frac{1}{\tau + 1} \right),$$

and the third by claim (ii). The last two inequalities also use the convexity of V_N with respect to the mean of a normal prior, i.e., Proposition 2. (Although Proposition 2 assumes normal priors for both arms, this can be relaxed.) It follows from (31) that

$$V_N^1(f_1; f_2; A_n) \leq V_N^1(\tilde{f}_1; f_2; A_n).$$

The rest of the proof is standard. □

8 Discussion

Results in previous sections suggest the following conjecture. Consider a two-armed bandit in the general exponential family setting with conjugate priors. Suppose the prior expected yield of one pull from each arm is the same, but the prior weight of arm 1 is larger. Then it seems reasonable that arm 2 is optimal at the first stage, i.e., in the notation of Section 3,

$$\frac{\gamma_1}{\tau_1} = \frac{\gamma_2}{\tau_2} \quad \text{and} \quad \tau_1 > \tau_2 \quad \implies \quad \Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \leq 0.$$

This holds if the discount sequence is infinite-horizon geometric. Indeed, it is optimal to pull arm 2 because, according to Corollary 2, arm 2 has a larger Gittins index. For non-geometric discounting, we cannot apply Corollary 2 due to the lack of an index policy. In fact, Berry (1972) proposed this conjecture for Bernoulli bandits with uniform discounting, and this special case is still open.

References

- [1] D. A. Berry, A Bernoulli two-armed bandit, *Ann. Math. Statist.* **43** (1972) 871–897.
- [2] D. A. Berry and B. Fristedt, Bernoulli one-armed bandits—arbitrary discount sequences, *Ann. Statist.* **7** (1979) 1086–1105.
- [3] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments* (1985) Chapman and Hall, New York.
- [4] R. N. Bradt, S. M. Johnson and S. Karlin, On sequential designs for maximizing the sum of n observations, *Ann. Math. Statist.* **27** (1956) 1060–1074.
- [5] L. D. Brown, *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory* (1986) Institute of Mathematical Statistics, Hayworth, CA.
- [6] M. K. Chattopadhyay, Two-armed Dirichlet bandits with discounting, *Ann. Statist.* **22** (1994) 1212–1221.
- [7] H. Chernoff, Optimal stochastic control, *Sankhya A* **30** (1968) 221–252.
- [8] H. Chernoff and A. J. Petkau, Numerical solutions for Bayes sequential decision problems, *SIAM J. Scient. Comput.* **7** (1986) 46–59.
- [9] M. K. Clayton and D. A. Berry, Bayesian nonparametric bandits, *Ann. Statist.* **13** (1985) 1523–1534.
- [10] J. C. Gittins, Bandit processes and dynamic allocation indices (with discussion), *Journal of the Royal Statistical Society, Series B* **41** (1979) 148–177.
- [11] J. C. Gittins and D. M. Jones, A dynamic allocation index for the sequential design of experiments. In: J. Gani, Editor, *Progress in Statistics*, North-Holland, Amsterdam (1974) 241–266.
- [12] J. C. Gittins and Y.-G. Wang, The learning component of dynamic allocation indices, *Ann. Statist.* **20** (1992) 1625–1636.
- [13] S. J. Herschkorn, Bandit bounds from stochastic variability extrema, *Stat. Prob. Lett.* **35** (1997) 283–288.
- [14] S. Karlin, *Total Positivity*, Stanford Univ. Press (1968).

- [15] H. Kaspi and A. Mandelbaum, Multi-armed bandits in discrete and continuous time, *Ann. Appl. Probab.* **8** (1998) 1270–1290.
- [16] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York (1979).
- [17] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, Wiley & Sons, Chichester (2002).
- [18] U. Rieder and H. Wagner, Structured policies in the sequential design of experiments, *Annals of Operations Research* **32** (1991) 165–188.
- [19] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*, Springer, New York (2007).
- [20] W. Whitt, Uniform conditional variability ordering of probability distributions, *Journal of Applied Probability* **22** (1985) 619–633.
- [21] P. Whittle, Multi-armed bandits and the Gittins index, *J. Roy. Statist. Soc. B* **42** (1980) 143–149.
- [22] Y.-C. Yao, Some results on the Gittins index for a normal reward process, in H.-C. Ho, C.-K. Ing, T. L. Lai, eds., *Time Series and Related Topics: In Memory of Ching-Zong Wei*, Institute of Mathematical Statistics, Beachwood, Ohio (2006) 284–294.
- [23] Y. Yu, On the entropy of compound distributions on nonnegative integers, *IEEE Transactions on Information Theory* **55** (2009a) 3645–3650.
- [24] Y. Yu, Monotonic convergence in an information theoretic law of small numbers, *IEEE Transactions on Information Theory* **55** (2009b) 5412–5422.
- [25] Y. Yu, Relative log-concavity and a pair of triangle inequalities, *Bernoulli* **16** (2010) 459–470.
- [26] Y. Yu, Prior ordering and monotonicity in Dirichlet bandits, *Preprint* arXiv:1101.4903 (2011).